Using Stochastic Models to Describe and Predict Social Dynamics of Web Users

KRISTINA LERMAN, USC Information Sciences Institute
TAD HOGG, Institute for Molecular Manufacturing

The popularity of content in social media is unequally distributed, with some items receiving a disproportionate share of attention from users. Predicting which newly-submitted items will become popular is critically important for both the hosts of social media content and its consumers. Accurate and timely prediction would enable hosts to maximize revenue through differential pricing for access to content or ad placement. Prediction would also give consumers an important tool for filtering the content. Predicting the popularity of content in social media is challenging due to the complex interactions between content quality and how the social media site highlights its content. Moreover, most social media sites selectively present content that has been highly rated by similar users, whose similarity is indicated implicitly by their behavior or explicitly by links in a social network. While these factors make it difficult to predict popularity *a priori*, stochastic models of user behavior on these sites can allow predicting popularity based on early user reactions to new content. By incorporating the various mechanisms through which web sites display content, such models improve on predictions that are based on simply extrapolating from the early votes. Specifically, for one such site, the news aggregator Digg, we show how a stochastic model distinguishes the effect of the increased visibility due to the network from how interested users are in the content. We find a wide range of interest, distinguishing stories primarily of interest to users in the network ("niche interests") from those of more general interest to the user community. This distinction is useful for predicting a story's eventual popularity from users' early reactions to the story.

Categories and Subject Descriptors: H.4.3 [**Information Systems Applications**]: Communications Applications

General Terms: Design, Measurement, Human Factors

Additional Key Words and Phrases: Social media, modeling, prediction, social news, social dynamics, social networks

ACM Reference Format:
Lerman, K. and Hogg, T. 2012. Using stochastic models to describe and predict social dynamics of web users. ACM Trans. Intell. Syst. Technol. 3, 4, Article 62 (September 2012), 33 pages.
DOI = 10.1145/2337542.2337547 http://doi.acm.org/10.1145/2337542.2337547

1. INTRODUCTION

Success or popularity in social media is not evenly distributed. Instead, a small number of users dominate activity on the site and receive most of the attention of other

This article is based on work supported in part by the Air Force Office for Scientific Research under contract FA9550-10-1-0102, by the National Science Foundation under grant IIS-0968370, and by the Defense Advanced Research Projects Agency under contract W911NF-12-1-0034. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

Authors' addresses: K. Lerman, Information Sciences Institute, University of Southern California; T. Hogg, Institute for Molecular Manufacturing.

DOI 10.1145/2337542.2337547 http://doi.acm.org/10.1145/2337542.2337547

users. The popularity of contributed items likewise shows extreme diversity. For example, relatively few of the four billion images on the social photo-sharing site Flickr are viewed thousands of times, while most of the rest are rarely viewed. Of the tens of thousands of new stories submitted daily to the social news portal Digg, only a handful become wildly popular, gathering thousands of votes, while most of the remaining stories never receive more than a single vote from the submitter herself. Among thousands of new blog posts every day, only a handful become widely read and commented upon. Given the volume of new content, it is critically important to provide users with tools to help them sift through the vast stream of new content to identify interesting items in a timely manner, or at least those items that will prove to be successful or popular. Accurate and timely prediction will also enable social media companies that host user-generated content to maximize revenue through differential pricing for access to content or ad placement, and encourage greater user loyalty by helping their users quickly find interesting new content.

Success in social media is difficult to predict. Although early and late popularity, which can be measured in terms of user interest (e.g., votes or views) an item generates from its inception are somewhat correlated [Gómez et al. 2008; Szabo and Huberman 2010], we know little about what drives success. Does success derive mainly from an item's inherent quality [Agarwal et al. 2008], users' response to it [Crane and Sornette 2008], or some external factors, such as social influence [Lerman 2007b; Lerman and Galstyan 2008; Lerman and Jones 2007]? In a landmark study, Salganik et al. [2006] addressed this question experimentally by measuring the impact of content quality and social influence on the eventual popularity or success of cultural artifacts. They showed that while quality contributes only weakly to eventual success, social influence, or knowing about the choices of other people, is responsible for both the inequality and unpredictability of success. In their experiment, Salganik et al. asked users to rate songs they listened to. The users were assigned to different groups. In the control group (independent condition), users were simply presented with lists of songs. In the other group (social influence condition), users were also shown how many times each song was downloaded by other users. The social influence condition resulted in a large inequality in popularity, measured by the number of times the songs were downloaded. While a song's quality, as measured by its popularity in the control group, was positively related to its eventual popularity in the social condition group, the variance in popularity at a given quality was very high. Thus two songs of similar quality could end up with vastly different levels of success. Moreover, when users were aware of the choices made by others, popularity was also unpredictable, meaning that on repeating the experiment, the same song could end up with a very different level of popularity.

Although Salganik et al.'s study was limited to a small set of songs created by unknown bands, its conclusions about the inequality and unpredictability of popularity appear to apply to cultural artifacts in general and social media production in particular. While this would appear to prohibit prediction of popularity, we argue that understanding how the collective behavior of Web users emerges from the decisions made by interconnected individuals allows predicting eventual popularity of items from the users' early reactions to them. As in previous work [Hogg and Lerman 2009; Hogg and Szabo 2009; Lerman 2007a], we use a stochastic modeling framework to mathematically describe the social dynamics of Web users. With this approach we studied the social news aggregator Digg. We produced a model that helps explain and predict [Lerman and Hogg 2010] the social voting patterns on Digg and related these aggregate behaviors to how Digg enables users to discover new content.

This prior work included social influence, that is, the increased visibility of stories to a user's neighbors in the social network, but did not address the commonality of users' interests indicated by links in the social network. This phenomenon, known as *homophily*, is a key aspect of social networks. In this article we present a new extension to the model that accounts for homophily by incorporating systematic variations of interests within and outside of the network neighborhood. The new model also more closely matches Web site behavior than previous studies. First, the new model's parameters account for the daily variation in user activity [Szabo and Huberman 2010], thereby focusing on how much votes received by individual stories deviate from the average activity rate on the site. Second, the model allows for the variation in the number of votes a story receives before it is promoted, which the prior model ignored.

By separating the impact of story quality and social influence on the popularity of stories on Digg, a stochastic model of social dynamics enables two novel applications: (1) estimating inherent story quality from the evolution of its observed popularity; and (2) predicting its eventual popularity based on users' early reactions to the story. Specifically, to predict how popular a story will become, we use the early votes, even those cast before the story is promoted, to estimate how interesting it is to the user community. With this estimate, the model then determines, on average, the story's subsequent evolution. We study these claims empirically on a sample of stories from Digg. We show that adjusting for the differing interests among voters based upon the social network improves predictions of popularity from the early reactions of users.

The article is organized as follows. In Section 2 we describe the social news aggregator Digg, which provides an empirical foundation and a dataset for investigating the utility of stochastic models on the prediction task. Section 3 presents an overview of the stochastic modeling framework. In Section 4 we apply the framework to study dynamics of social voting on Digg. We review a prior model of social dynamics of Digg and show that it explains many of the empirically observed features of aggregate behavior of voters on that site. In Section 5 we extend this model to include variations in story interest to users based on their links in the social network, to account for homophily. Then, in Section 6 we show how the model can predict eventual popularity of newly submitted stories on Digg.

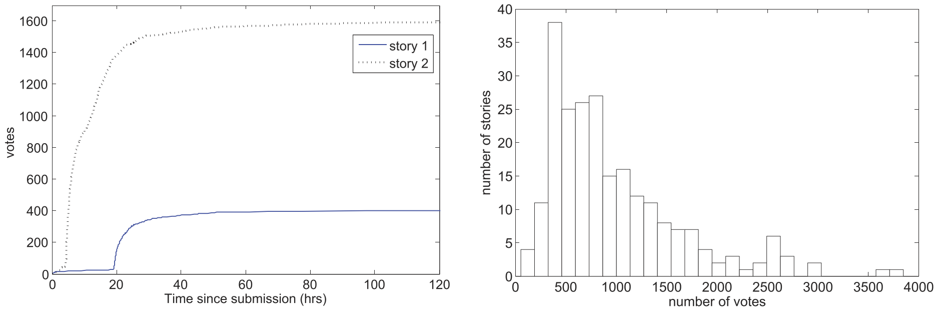
2. SOCIAL NEWS PORTAL DIGG

With over three million registered users, the social news aggregator Digg is one of the more popular news portals on the Web. Digg allows users to submit and rate news stories by voting on, or "digging," them. There are many new submissions every minute, over 16,000 a day. Every day Digg picks about a hundred stories that it believes will be most interesting to the community and promotes them to the front page. Although the exact promotion mechanism is kept secret and changes occasionally, it appears to take into account the number of votes the story receives and how rapidly it receives them. Digg's success is fueled in large part by the emergent front page, which is created by the collective decision of its many users.

While the life cycle of each story may be drastically different from others, its basic elements are the same. These are specified by Digg's user interface, which defines how users post or discover new stories and interact with other users. A model of social dynamics has to take these elements into account when describing the evolution of a story's popularity.

2.1. User Interface

A newly submitted story goes on the *upcoming* stories list, where it remains for a period of time, typically 24 hours, or until it is promoted to the front page, whichever comes first. The default view shows newly submitted stories as a chronologically ordered list, with the most recently submitted story at the top of the list, 15 stories



(a) (b)

Fig. 1. Dynamics of social voting: (a) evolution of the number of votes received by two front page stories in June 2006; (b) distribution of popularity of 201 front page stories submitted in June 2006.

to a page. To see older stories, a user must navigate to page 2, 3, and so on, of the upcoming stories list. Promoted stories (Digg calls them "popular") are also displayed as a chronologically ordered list on the *front pages*, 15 stories to a page, with the most recently promoted story at the top of the list. To see older promoted stories, the user must navigate to pages 2, 3, and so on, of the front page. Users vote for the stories they like by "digging" them. The yellow badge to the left of each story shows its current popularity.

Digg allows users to designate friends and track their activities, that is, see the stories that friends recently submitted or voted for. The *friends interface* is available through the "Friends' Activity" link at the top of any Digg Web page. The friend relationship is asymmetric. When user A lists user B as a *friend*, A can watch the activities of B but not vice versa. We call A the *fan* of B. A newly submitted story is visible in the upcoming stories list, as well as to the submitter's fans via the friends interface. With each vote, a story becomes visible to the voter's fans through the friends interface, which shows the newly submitted stories that the user's friends voted for.

Digg allows users to view the most popular stories from the previous day, week, month, or year. Digg also implements a social filtering feature which recommends stories, including upcoming stories, that were liked by users with a similar voting history. This interface, however, was not available at the time the data for our study was collected, and hence is not part of the stochastic models described in this article. Thus we examine a period of time when Digg had a relatively simple user interface, which simplifies the stochastic models.

2.2. Dynamics of Popularity

While a story is in the upcoming stories list, it accrues votes slowly. If the story is promoted to the front page, it accumulates votes at a much faster pace. Figure 1(a) shows the evolution of the number of votes for two stories submitted in June 2006. The point where the slope abruptly increases corresponds to promotion to the front page. The vast majority of stories are never promoted and, therefore, never experience the sharp rise in the number of votes that accompanies being featured on the front page. As the story ages, accumulation of new votes slows down [Wu and Huberman 2007], and after a few days the total number of votes received by a story saturates to some value. This value, which we call the final number of votes, gives a measure of the story's success or *popularity*.

Popularity varies widely from story to story. Figure 1(b) shows the distribution of the final number of votes received by front page stories that were submitted over a period of about two days in June 2006. The distribution shows "inequality of popularity":

a handful of stories become very popular, accumulating thousands of votes, while most others only muster a few hundred votes. This distribution applies to front page stories only. Stories that are never promoted to the front page receive very few votes, in many cases just a single vote from the submitter. In systems displaying such "long tailed" distributions, extreme events, for example, a story receiving many thousands of votes, occur much more frequently than would be expected if the underlying processes were Poisson or Gaussian in nature.

Long tails are ubiquitous features of human activity [Anderson 2006]. Examples include inequality of popularity of cultural artifacts, such as books and music albums [Salganik et al. 2006], and in a variety of online behaviors [Wilkinson 2008], including tagging, where a few documents are tagged much more frequently than others, collaborative editing on Wikis [Kittur et al. 2006], and votes on a sample of more than 30,000 stories promoted to Digg's front page over the course of a year [Wu and Huberman 2007].

While unpredictability of popularity is more difficult to verify than in the controlled experiments of Salganik et al., it is reasonable to assume that a similar set of stories submitted to Digg on another day will end with radically different numbers of votes. In other words, while the distribution of the final number of votes these stories receive will look similar to the distribution in Figure 1(b), the number of votes received by individual stories will be very different in the two realizations.

2.3. Data Collection

We collected data for the study by scraping Digg's Web pages in May and June 2006. The May dataset consists of stories that were submitted to Digg May 25 to 27, 2006. We followed these stories by periodically scraping Digg to determine the number of votes stories received as a function of the time since their submission. We collected at least four such observations for each of 2152 stories, submitted by 1212 distinct users. Of these stories, 510, by 239 distinct users, were promoted to the front page. We followed the promoted stories over a period of several days, recording the number of votes the stories received. This May dataset also records the location of the stories on the upcoming and front pages as a function of time.

The June dataset consists of 201 stories promoted to the front page between June 27 and 30, 2006. For each story, we collected the names of its first 216 voters.

We focus on the early stages of story evolution – from submission until shortly after promotion – because the Digg social network has a much larger effect on upcoming than front page stories due to the much more rapid addition of stories to the upcoming list. This large influx of stories makes it difficult for users to find a new story before it becomes hidden by the arrival of more stories. In this case, enhanced visibility via the network for fans of the submitter or early voters is particularly important, and a model of social dynamics has to account for it. In light of these observations, and for speeding up data collection, we focus on the early votes for stories.

Activity on Digg varies considerably over the course of a day, as seen in Figure 2. Adjusting times by the cumulative activity on the site accounts for this variation and improves predictions [Szabo and Huberman 2010]. We define the "Digg time" between two events (e.g., votes on a story) as the total number of votes on front page stories during the time between those events. This behavior is similar to that seen in an extensive study of front page activity in 2007 [Szabo and Huberman 2010], and as in that study, we scale the measure by defining a "Digg hour" to be the average number of front page votes in an hour, which is 2500 for our dataset. We evaluate the consequence of this variability by contrasting a model based on real time (in Section 4) with one based on Digg time (in Section 5).

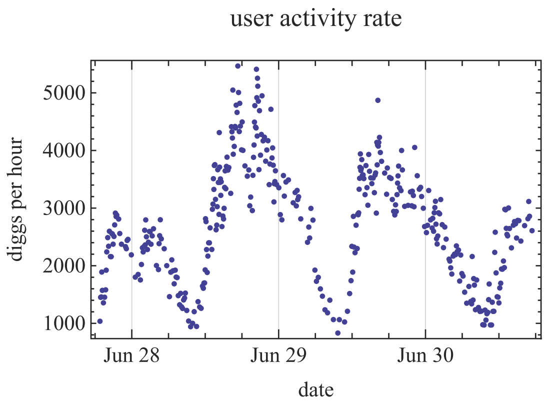


Fig. 2. Voting rate (diggs per hour) on front page stories at the end of June 2006. The dates indicated are the start of each day (0:00 GMT). The minimum in daily activity is around 9am GMT. Each point is the average voting rate for 100 successive votes.

In addition to voter activity, we extracted a snapshot of the social network of the top-ranked 1020 Digg users as of June 2006. This data contained the names of each user's friends and fans. Since the original network did not contain information about all the voters in our data, we augmented it in February 2008 by extracting the names of friends of about 15, 000 additional users. Many of these users added friends between June 2006 and February 2008. Although Digg does not provide the time a new link was created, it lists the links in reverse chronological order and gives the date the friend joined Digg. By eliminating friends who joined Digg after June 30, 2006, we were able to reconstruct the fan links for all voters in our data. This data allows us to identify, for each vote, whether the user was a fan of any prior voter on that story, in which case the story would have appeared in the friends interface for that user.

Votes by fans account for 6% of the votes in the June dataset and about 3% of the front page votes.

The datasets used in this and previous work were collected before Digg's API was introduced. Scraping Web pages to extract data had several issues. First, data had to be manually cleaned to ensure consistency. Second, since vote timestamps were not available on the Web page, we had to supplement June 2006 data by using the Digg API in October 2009 to obtain the time of each vote, the final number of votes the story received, and the time of promotion. In the intervening time, however, some of the users had deleted their accounts. Since we could not easily resolve the time of the vote of an inactive user, we had to delete these users from the voters list. We believe that the small fraction of data lost in this manner (less than 8% of the data) does not adversely affect the modeling study. However, in the future we plan to repeat the study on a much cleaner dataset obtained through the Digg API.

3. STOCHASTIC MODELS OF SOCIAL DYNAMICS

Rather than account for the inherent variability of individuals, stochastic models focus on the macroscopic, or aggregate, behavior of the system, which can be described by average quantities. In the context of Digg, such quantities include the average rate at which users post new stories and vote on existing stories. Such macroscopic descriptions often have a simple form and are analytically tractable. Stochastic models do not reproduce the results of a single observation—rather, they describe the typical behavior. These models are analogous to the approach used in statistical physics, demographics, and macroeconomics where the focus is on relations among aggregate quantities, such as volume and pressure of a gas, population of a country and immigration, or interest rates and employment.

We represent each individual entity, whether a user or a story, as a stochastic process with a few states. This abstraction captures much of the individual complexity and environmental variability by casting an individual's actions as inducing probabilistic transitions between states. While this modeling framework applies to stochastic processes of varying complexity, for simplicity, we focus on processes that obey the Markov property, namely, a user whose future state depends only on her present state and the input she receives. A Markov process can be succinctly captured by a *state diagram* showing the possible states of the user and conditions for transition between those states. This approach is similar to compartmental models in biology [Ellner and Guckenheimer 2006]. For instance, in epidemiology such models track the progress of a disease as shifting individuals between states, or compartments, such as susceptible and infected.

We assume that all users have the same set of states, and that transitions between states depend only on the state and not the individual user. That is, the state captures the key relevant properties determining subsequent user actions. A choice of states to describe users results in grouping users in the same state into the same compartment for modeling. Then, the aggregate state of the system can be described simply by the number of individuals in each state at a given time. That is, the system configuration at this time is defined by the occupation vector: $\vec{n} = (n_1, n_2, \ldots)$ where n_k is the number of individuals in state k.

We focus on modeling the behavior (i.e., votes received) of individual stories. Thus in our application of this approach, there is a different occupation vector for each story. For example, the states of a user with respect to a given story on Digg could be "has not seen the story," "has seen the story but did not vote for it," and "has voted for the story". The corresponding occupation vector has three elements, counting the number of users in each of these three compartments at a given time. As the story gains votes, users transition to the "has voted for the story" state, increasing the value of the corresponding element of the occupation vector. As described below, in our application of this approach to social media, we include the social network links of the users as part of the state, and hence the occupation vectors we use have more than three elements.

A key requirement for designing stochastic models is to ensure the state captures enough of the large variation in individual behavior to give a useful description of aggregate system properties. This is particularly challenging when individual activity follows a long-tail distribution, such as seen in some epidemics [Lloyd-Smith et al. 2005], as well as in social media Web sites [Barabási 2005; Wilkinson 2008]. In our case, including user link information as part of the state accounts for enough of this variation to provide reasonable accuracy—in particular, significantly improving predictions compared to direct extrapolation of voting rates without accounting for the properties of the Web site user interface.

The next step in developing the stochastic model is to summarize the variation within the collection of histories of changing occupation vectors with a probabilistic description. That is, we characterize the possible occupation vectors by the probability, $P(\vec{n}, t)$, the system is in configuration $\vec{n}$ at time t. The evolution of $P(\vec{n}, t)$, governed by the Stochastic Master Equation [Kampen 1992], is almost always too complex to be analytically tractable. Fortunately, we can simplify the problem by working with the average occupation number, whose evolution is given by the rate equation:

$$\frac{d\langle n_k \rangle}{dt} = \sum_j w_{jk}(\langle \vec{n} \rangle)\langle n_j \rangle - \langle n_k \rangle \sum_j w_{kj}(\langle \vec{n} \rangle) \qquad (1)$$

where $\langle n_k \rangle$ denotes the average number of users in state k at time t, that is, $\sum_{\vec{n}} n_k P(\vec{n}, t)$ and $w_{jk}(\langle \vec{n} \rangle)$ is the transition rate from configuration j to configuration k when the occupation vector is $\langle \vec{n} \rangle$.

Using the average of the occupation vector in the transition rates is a common simplifying technique for stochastic models. A sufficient condition for the accuracy of this approximation is that variations around the average are relatively small. In many stochastic models of systems with large numbers of components, variations are indeed small due to many independent interactions among the components and the short tails of the distributions of these component behaviors. More elaborate versions of the stochastic approach give improved approximations when variations are not small, particularly due to correlated interactions [Opper and Saad 2001] or large individual heterogeneity [Moreno et al. 2002]. User behavior on the Web, however, often involves distributions with long tails, whose typical behaviors differ significantly from the average [Barabási 2005; Wilkinson 2008]. In this case we have no guarantee that the averaged approximation is adequate, even when aggregating the behavior of many users [Sornette 2004]. Instead, we must test its accuracy for particular aggregate behaviors by comparing model predictions with observations of actual behavior, as we report below.

In the rate equation, occupation number n_k increases due to users' transitions from other states to state k, and decreases due to transitions from the state k to other states. The equations can be easily written down from the user state diagram. Each state corresponds to a dynamic variable in the mathematical model—the average number of users in that state—and it is coupled to other variables via transitions between states. Every transition must be accounted for by a term in the equation, with transition rates specified by the details of the interactions between users.

In summary, the stochastic modeling framework is quite general and requires only specifying the aggregate states of interest for describing the system and how individual user behaviors create transitions among these states. The modeling approach is best suited to cases where the users' decisions are mainly determined by a few characteristics of the user and the information they have about the system. These system states and transitions give the rate equations. Solutions to these equations then give estimates of how aggregate behavior varies in time and depends on the characteristics of the users involved.

4. A MODEL OF SOCIAL DYNAMICS OF DIGG

Underlying a stochastic model of social dynamics is a behavioral model of an individual Web user. The behavioral model accounts for choices that a Web site's user interface allows users. Detailed data about human activity that can be collected from social media sites such as Digg allow us to parameterize the models and test them by comparing their predictions to the observed collective dynamics.

A prior study of social dynamics of Digg [Hogg and Lerman 2009] used a simple behavioral model that viewed each Digg user as a stochastic Markov process, whose state diagram with respect to a single story is shown in Figure 3. According to this model, a user visiting Digg can choose to browse the *front* pages to see the recently promoted stories, *upcoming* stories pages for the recently submitted stories, or use the *friends* interface to see the stories her friends have recently submitted or voted for. She can select a story to read from one of these pages and, if she considers it interesting, vote for it. The user's environment, the stories she is seeing, changes in time due to the actions of all the users.

We characterize the changing state of a story by three values: the number of votes, $N_{\mathrm{vote}}(t)$ that the story has received by time t after it was submitted to Digg; the list the

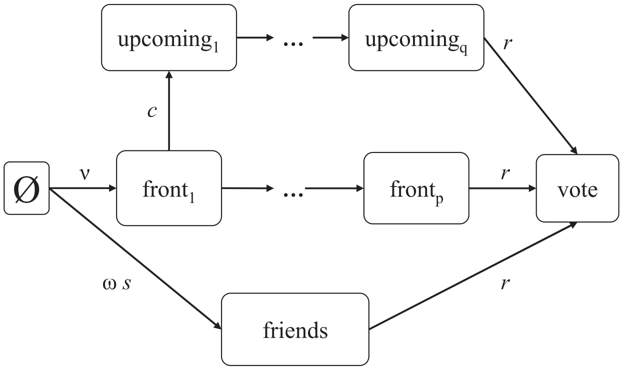


Fig. 3. State diagram of user behavior for a single story. A user starts in the Ø state at the left, may find the story through one of the three interfaces and may then vote on it. At a given time, the story is located on a particular page of either the upcoming or front page lists, not both. This diagram shows votes for a story on either page p of the front pages or page q of the upcoming pages. Only fans of previous voters can see the story through the friends interface. Users in the friends, front or upcoming states may choose to leave Digg, thereby returning to the Ø state (with those transitions not shown in the figure). Users reaching the "vote" state remain there indefinitely and cannot vote on the story again. Parameters next to the arrows characterize state transitions.

story is in at time t (*upcoming* or *front* page); and its location within that list, which we denote by q and p for upcoming and front page lists, respectively.

With Figure 3 as a modeling blueprint, we relate the users' choices to the changes in the state of a single story. In terms of the general rate equation (Eq. (1)), the occupancy vector $\vec{n}$ describing the aggregate user behavior at a given time has the following components: the number of users who see a story via one of the front pages; one of the upcoming pages, through the friends pages; and number of users who vote for a story, N_{vote}. Since we are interested in the number of users who reach the vote state, we do not need a separate equation for each state in Figure 3: at a given time, a particular story has a unique location on the upcoming or front page lists. Thus, for simplicity, we can group the separate states for each list in Figure 3, and consider just the combined transition for a user to reach the page containing the story at the time she visits Digg. These combined transition rates depend on the location of the story in the list, that is, the value of q or p for the story. With this grouping of user states, the rate equation for $N_{\text{vote}}(t)$ is

$$\frac{dN_{\text{vote}}(t)}{dt} = r(\nu_{\text{f}}(t) + \nu_{\text{u}}(t) + \nu_{\text{friends}}(t)) \tag{2}$$

where r measures how interesting the story is, that is, the probability a user seeing the story will vote on it, and ν_{f}, ν_{u} and ν_{friends} are the rates at which users find the story via one of the front or upcoming pages, and through the friends interface, respectively.

In this model, the transition rates appearing in the rate equation depend on the time t, but not on the occupation vector. Nevertheless, the model could be generalized to include such a dependence if, for example, a user currently viewing an interesting story not only votes on it but explicitly encourages people they know to view the story as well.

4.1. Story Visibility

Before we can solve Eq. (2), we must model the rates at which users find the story through the various Digg interfaces. These rates depend on the story's location in the list. The parameters of these models depend on user behaviors that are not readily measurable. Instead, we estimate them using data collected from Digg, as described below.

Visibility by position in the list. A story's visibility on the front page or upcoming stories lists decreases as recently added stories push it further down the list. The stories are shown in groups: the first page of each list displays the 15 most recent stories, page 2 the next 15 stories, and so on.

We lack data on how many Digg visitors proceed to pages 2, 3, and so on, in each list. However, when presented with lists over multiple pages on a Web site, successively smaller fractions of users visit later pages in the list. One model of users following links through a Web site considers users estimating the value of continuing at the site, and leaving when that value becomes negative [Huberman et al. 1998]. This model leads to an inverse Gaussian distribution of the number of pages m a user visits before leaving the Web site,

$$e^{-\frac{\lambda(m-\mu)^2}{2m\mu^2}}\sqrt{\frac{\lambda}{2\pi m^3}} \tag{3}$$

with mean μ and variance μ^3/λ. This distribution matches empirical observations in several Web settings [Huberman et al. 1998]. When the variance is small, for intermediate values of m this distribution approximately follows a power law, with the fraction of users leaving after viewing m pages decreasing as $m^{-3/2}$.

To model the visibility of a story on the mth front or upcoming page, the relevant distribution is the fraction of users who visit *at least* m pages, that is, the upper cumulative distribution of Eq. (3). For $m > 1$, this fraction is

$$f_{\text{page}}(m) = \frac{1}{2}\left(F_m(-\mu) - e^{2\lambda/\mu}F_m(\mu)\right) \tag{4}$$

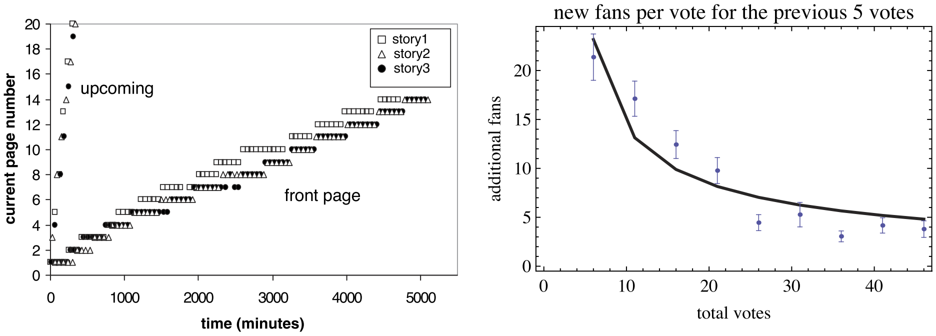
where $F_m(x) = \text{erfc}(\alpha_m(m-1+x)/\mu)$, erfc is the complementary error function, and $\alpha_m = \sqrt{\lambda/(2(m-1))}$. For $m = 1$, $f_{\text{page}}(1) = 1$.

The visibility of stories decreases in two distinct ways when a new story arrives. First, a story moves down the list on its current page. Second, a story at the 15th position moves to the top of the next page. For simplicity, we model these processes as decreasing visibility, that is, the value of $f_{\text{page}}(m)$, through m taking on fractional values within a page, that is, $m = 1.5$ denotes the position of a story half way down the list on the first page. This model is likely to somewhat overestimate the loss of visibility for stories among the first few of the 15 items on a given page, since the top several stories are visible without requiring the user to scroll down the page.

List position of a story. Figure 4(a) shows how the page number of a story on the two lists changes in time for three randomly chosen stories from our dataset. The behavior is close to linear when averaging over the daily activity variation (shown in Figure 2). For simplicity in this model, we ignore this variation and take a story's page number on the upcoming page q and the front page p at time t to be [Hogg and Lerman 2009]:

$$p(t) = k_{\text{f}}(t - T_{\text{promotion}}) + 1 \tag{5}$$
$$q(t) = k_{\text{u}}t + 1 \tag{6}$$

where $T_{\text{promotion}}$ is the time the story is promoted to the front page (or ∞ if the story is never promoted) and the slopes are given in Table I. For a given story, $p(t)$ is only defined for times $t \geq T_{\text{promotion}}$ and $q(t)$ for $t < T_{\text{promotion}}$. Since each page holds 15 stories, these rates are 1/15th the submission and promotion rates, respectively.



(a) (b)

Fig. 4. (a) Current page number on the upcoming and front pages vs. time for three different stories. Time is measured from when the story first appeared on each page, that is, time it was submitted or promoted, for the upcoming and front page points, respectively. (b) Increase in the number of distinct users who can see the story through the friends interface with each group of five new votes for the first 46 users to vote on a story. The points are mean values for 195 stories, including those shown in (a), and the curve is based on Eq. (7). The error bars indicate the standard error of the estimated means.

Table I. Model Parameters

parameter	value
rate general users come to Digg	$\nu = 600$ users/hr
fraction viewing upcoming pages	$c = 0.3$
rate a voters' fans come to Digg	$\omega = 0.12/hr$
page view distribution	$\mu = 0.6$, $\lambda = 0.6$
fans per new vote	$a = 51$, $b = 0.62$
vote promotion threshold	$h = 40$
upcoming stories location	$k_{\mathrm{u}} = 3.60$ pages/hr
front page location	$k_{\mathrm{f}} = 0.18$ pages/hr
story specific parameters	
interestingness	r
number of submitter's fans	S

Front page and upcoming stories lists. Digg prominently shows the stories on the front page. The upcoming stories list is less popular than the front page. We model this fact by assuming a fraction $c < 1$ of Digg visitors proceed to the pages of upcoming stories.

We use a simple threshold to model how a story is promoted to the front page. Initially, the story is visible on the upcoming stories pages. If and when the number of votes a story receives exceeds a promotion threshold h, the story moves to the front page. This threshold model approximates Digg's promotion algorithm as of May 2006, since in our dataset we did not see any front page stories with fewer than 44 votes, nor did we see any upcoming stories with more than 42 votes. We take $h = 40$ as an approximation to the promotion algorithm.

Friends interface. The friends interface allows the user to see the stories her friends have (i) submitted, (ii) voted for, and (iii) commented on in the preceding 48 hours. Although users can take advantage of all these features, we only consider the first two. These uses of the friends interface are similar to the functionality offered by other social media sites: for example, Flickr allows users to see the latest images his friends uploaded, as well as the images a friend liked.

The fans of the story's submitter can find the story via the friends interface. As additional people vote on the story, their fans can also see the story. We model this with $s(t)$, the number of fans of voters on the story by time t who have not yet seen the story. Although the number of fans is highly variable, the average number of additional fans from an extra vote when the story has N_{vote} votes is approximately

$$\Delta s = aN_{\text{vote}}^{-b} \tag{7}$$

where $a = 51$ and $b = 0.62$, as illustrated in Figure 4(b), showing the fit to the increment in average number of fans per vote over groups of 5 votes as given in the data. Thus early voters on a story tend to have more new fans (i.e., fans who are not also fans of earlier voters) than later voters.

The model can incorporate any distribution for the times fans visit Digg. We suppose these users visit Digg daily, and since they are likely to be geographically distributed across all time zones, the rate at which fans discover the story is distributed throughout the day. A simple model of this behavior takes fans arriving at the friends page independently at a rate ω. As fans read the story, the number of potential voters gets smaller, that is, s decreases at a rate ωs, corresponding to the rate fans find the story through the friends interface, ν_{friends}. We neglect additional reduction in s from fans finding the story without using the friends interface.

Combining the growth in the number of available fans and its decrease as fans return to Digg gives

$$\frac{ds}{dt} = -\omega s + aN_{\text{vote}}^{-b}\frac{dN_{\text{vote}}}{dt} \tag{8}$$

with initial value $s(0)$ equal to the number of fans of the story's submitter, S. This model of the friends interface treats the pool of fans uniformly. That is, we assume no difference in behavior, on average, for fans of the story's submitter vs. fans of other voters.

In summary, the rates in Eq. (2) are:[1]

$$\begin{aligned} \nu_{\text{f}} &= \nu f_{\text{page}}(p(t))\,\Theta(N_{\text{vote}}(t) - h) \\ \nu_{\text{u}} &= c\nu f_{\text{page}}(q(t))\,\Theta(h - N_{\text{vote}}(t))\Theta(24hr - t) \\ \nu_{\text{friends}} &= \omega s(t) \end{aligned}$$

where t is time since the story's submission and ν is the rate users visit Digg. The first step function in ν_{f} and ν_{u} indicates that when a story has fewer votes than required for promotion, it is visible in the upcoming stories pages; and when $N_{\text{vote}}(t) > h$, the story is visible on the front page. The second step function in ν_{u} accounts for a story staying in the upcoming list for at most 24 hours. We solve Eq. (2) subject to initial condition $N_{\text{vote}}(0) = 1$, because a newly submitted story starts with a single vote, from the submitter.

4.2. Model Parameters

The solutions of Eq. (2) show how the number of votes received by a story changes in time. The solutions depend on the model parameters, of which only two parameters, the story's interestingness r and number of fans the submitter has S, change from one story to another. Therefore, we fix values of the remaining parameters as given in Table I.

As just described, we estimate some of these parameters (such as the growth in list location, promotion threshold, and fans per new vote) directly from the data. The remaining parameters are not directly given by our dataset (e.g., how often users view the upcoming pages) and instead we estimate them based on the model predictions. The small number of stories in our dataset, as well as the approximations made in

[1] $\Theta(x)$ is a step function 1 when $x \geq 0$ and 0 when $x < 0$.

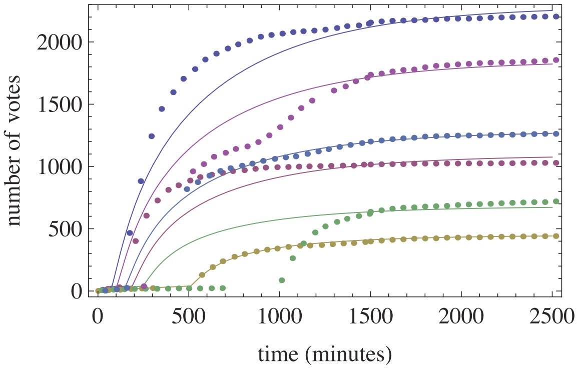


Fig. 5. Evolution of the number of votes received by six stories compared with model solution.

Table II. Parameters for the Example Stories (listed in decreasing order of total votes received by the story and hence corresponding to the curves in Figure 5 from top to bottom)

S	r	final votes
5	0.51	2229
5	0.44	1921
40	0.32	1297
40	0.28	1039
160	0.19	740
100	0.13	458

the model, do not give strong constraints on these parameters. We selected one set of values giving a reasonable match to our observations. For example, the rate fans visit Digg and view stories via the friend's interface, given by ω in Table I, has 90% of the fans of a new voter returning to Digg within the next 19 hours. As another example of interpreting these parameter values, for the page visit distribution the values of μ and λ in Table I correspond to about 1/6 of the users viewing more than just the first page. These parameters could in principle be measured independently from aggregate behavior with more detailed information on user behavior. Measuring these values for users of Digg, or other similar Web sites, could improve the choice of model parameters.

4.3. Results

The model describes the behavior of all stories, whether or not they are promoted to the front page. To illustrate the model results, we consider stories promoted to the front page. Figure 5 shows the behavior of six such stories. For each story, S is the number of fans of the story's submitter, available from our data, and r is estimated to minimize the root-mean-square (RMS) difference between the observed votes and the model predictions. Table II lists these values.

Overall there is qualitative agreement between the data and the model, indicating that the features of the Digg user interface we considered can explain the patterns of collective voting. Specifically, the model reproduces three generic behaviors of Digg stories: (1) slow initial growth in votes of upcoming stories; (2) more interesting stories (higher r) are promoted to the front page (inflection point in the curve) faster and receive more votes than less interesting stories; (3) however, as first described in Lerman

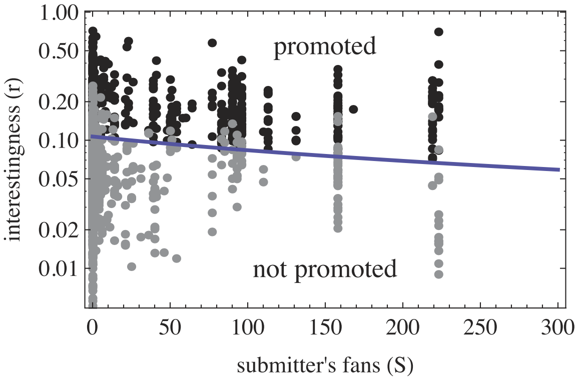


Fig. 6. Story promotion as a function of S and r. The r values are shown on a logarithmic scale. The model predicts stories above the curve are promoted to the front page. The points show the S and r values for the stories in our dataset: black and gray for stories promoted or not, respectively.

[2007b], better connected users (high S) are more successful in getting their less interesting stories (lower r) promoted to the front page than poorly-connected users. These observations highlight a benefit of the stochastic approach: identifying simple models of user behavior that are sufficient to produce the aggregate properties of interest.

The only significant difference between the data and the model is visible in the lower two lines of Figure 5. In the data, a story posted by the user with $S = 100$ fans is promoted before the story posted by the user with $S = 160$ fans, but saturates at smaller value of votes than the latter story. In the model, the story with larger r is promoted first and gets more votes.

Thus while the stochastic model is primarily intended to describe typical story behavior, we see it gives a reasonable match to the actual vote history of individual stories. Nevertheless, there are some cases where individual stories differ considerably from the model, particularly where an early voter happens to have an exceptionally large number of fans, thereby increasing the story's visibility to other users far more than expected. This variation, a consequence of the long-tail distributions involved in social media, is considerably larger than seen, for example, in most statistical physics applications of stochastic models. The effect of such large variations is an important issue to address when using stochastic models to predict the behavior of individual stories in social media.

Figure 6 shows parameters required for a story to reach the front page according to the model, and how that prediction compares to the stories in our dataset. The model's prediction of whether a story is promoted is correct for 95% of the stories in our dataset. For promoted stories, the correlation between S and r is -0.13, which is significantly different from zero (p-value less than 10^{-4} by a randomization test). Thus a story submitted by a poorly connected user (small S) tends to need high interest (large r) to be promoted to the front page [Lerman 2007b].

Figure 7 shows the estimated r values for the 510 promoted stories in our dataset have a wide range of interestingness to users. That is, even after accounting for the variation in visibility of the stories, there remains a significant range in how well stories appeal to users. Specifically, Figure 8 shows these r values fit well to a lognormal distribution

$$P_{\text{lognormal}}(\mu, \sigma; r) = \frac{1}{\sqrt{2\pi}\, r\sigma} \exp\left(-\frac{(\mu - \log(r))^2}{2\sigma^2}\right) \tag{9}$$

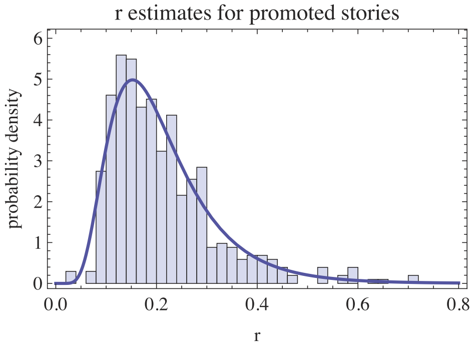


Fig. 7. Distribution of interestingness (i.e., *r* values) for the promoted stories in our dataset compared with the best fit lognormal distribution.

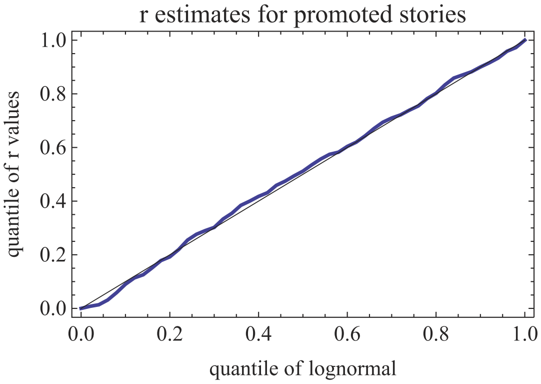


Fig. 8. Quantile-quantile plot comparing observed distribution of *r* values with the lognormal distribution fit (thick curve). For comparison, the thin straight line from 0 to 1 corresponds to a perfect match between the data and the distribution.

where parameters μ and σ are the mean and standard deviation of log(r). For the distribution of interestingness values, the maximum likelihood estimates of the mean and standard deviation of log(r) equal to -1.67 ± 0.04 and 0.47 ± 0.03, respectively, with the ranges giving the 95% confidence intervals. A randomization test based on the Kolmogorov–Smirnov statistic and accounting for the fact that the distribution parameters are determined from the data [Clauset et al. 2009] shows the *r* values are consistent with this distribution (p-value 0.35). While broad distributions occur in several Web sites [Wilkinson 2008], our model allows factoring out the effect of visibility due to the user interface from the overall distribution of votes. Thus we can identify variation in users' inclination to vote on a story they see.

The simple model described in this section gives a reasonable qualitative account of how user behavior leads to the promotion of stories to the front page and the eventual saturation in the number of votes they receive due to their decreasing visibility. In the section below we show how additional properties of the interface and user population can be added to the model for a more accurate analysis of the aggregate behavior. For example, a submitter's fans may find the story more interesting than the general Digg audience, corresponding to different *r* values for these groups of users. In addition, we modeled users coming to Digg independently with uniform rates ν and ω. In fact, the rates vary systematically over hours and days [Szabo and Huberman 2010] as shown in Figure 2, and individual users have a wide range in time between visits [Vázquez

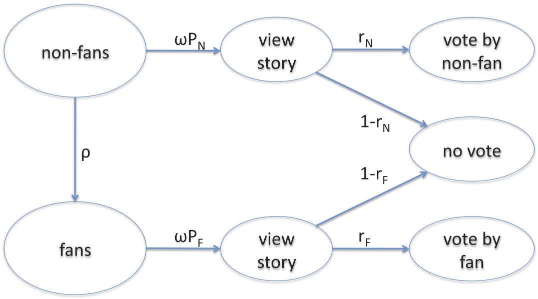


Fig. 9. State diagram for a user. The submitter provides a story's first vote. The initial set of fans consists of the submitter's fans; other users are initially nonfans. Fans and nonfans have different probabilities to see and vote on the story. With each vote, a nonfan user who is a fan of that voter moves into the *fans* state. This state transition is caused by the votes of other users: a user moving from the *nonfans* to *fans* state is not aware of that change until later visiting Digg and seeing the story in the friends interface.

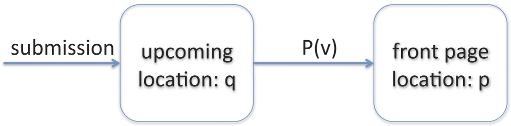


Fig. 10. State diagram for a story. A story starts at the top of the upcoming pages, with location $q = 1$. The location increases with each new submission. An upcoming story with v votes is promoted with probability $P(v)$. A promoted story starts at the top of the front pages, with location $p = 1$. The location increases as more stories are promoted. A story not promoted within a day is removed (not shown).

et al. 2006]. In our model, this variation gives time-dependent values for ν, describing the rate users come to Digg, and k_{f} and k_{u}, which relate to the rate new stories are posted and promoted.

The ability of the stochastic approach to incorporate additional details in the user models illustrates its value in providing insights into how aggregate behavior arises from the users, in contrast to models that evaluate regularities in the aggregate behaviors [Wu and Huberman 2007]. In particular, user models can help distinguish aggregate behaviors arising from intrinsic properties of the stories (e.g., their interestingness to the user population) from behavior due to the information the Web sites provides, such as ratings of other users and how stories are placed in the site, that is, visibility. Finally, stochastic models have not only explanatory, but also predictive power.

5. A MODEL OF SOCIAL VOTING WITH NICHE INTERESTS

To investigate differences among voters with respect to the fan network, we extend the previous stochastic model to distinguish votes from fans and nonfans, respectively, users who are, or are not, linked to previous voters through the fan network. The model considers the joint behavior of users and the location of the story on the Web site. Figure 9 shows the user states and the stochastic transitions between them. Stories are on either the upcoming or front pages, as shown in Figure 10. This leads to a description of the average rates of growth for votes from fans and nonfans of prior voters, v_F and v_N, respectively:

$$\frac{dv_F}{dt} = \omega r_F P_F F, \tag{10}$$

$$\frac{dv_N}{dt} = \omega r_N P_N N, \tag{11}$$

where t is the Digg time since the story's submission and ω is the average rate a user visits Digg (measured as a rate per unit Digg time). v_N includes the story's submitter; P_F and P_N denote the story's *visibility*; and r_F and r_N denote the story's *interestingness* to users who are fans or non-fans of prior voters, respectively. Visibility depends on the story's state (e.g., whether it has been promoted), as discussed below. Interestingness is the probability a user who sees the story will vote on it. Nominally people become fans of those whose contributions they consider interesting, suggesting fans likely have a systematically higher interest in stories. Our model accounts for this possibility with separate interestingness values for fans and nonfans.

In contrast to the model of Section 4 where time t denoted real time since story submission, we now use t to denote the "Digg time" since submission, thereby accounting for the daily variation in activity. As defined in Section 2.3, "Digg time" between two events (e.g., story submission and nth vote) measures the total number of votes on front page stories that took place between those events. Using Digg time reduces the variation in the rate users visit Digg, thereby improving the match to the assumed constant rate ω used in the model. Moreover, a detailed examination of the page locations of the stories in our dataset, shows systematic variation in the time stories spend on each page corresponding to the daily activity variation used to define Digg time. Thus using Digg time improves the accuracy of the linear growth in location given in Eqs. (5) and (6).

These voting rates depend on F (N), the numbers of users who have not yet seen the story and who are (are not) fans of prior voters. The quantities change as users see and vote on the story according to

$$\frac{dF}{dt} = -\omega P_F F + \rho N \frac{dv}{dt} \tag{12}$$

$$\frac{dN}{dt} = -\omega P_N N - \rho N \frac{dv}{dt} \tag{13}$$

with $v = v_F + v_N$ the total number of votes the story has received. The quantity ρ is the probability a user who has not yet seen the story and is not a fan of a prior voter is a fan of the most recent voter. For simplicity, we treat this probability as a constant over the voters, thus averaging over the variation due to clustering in the social network and the number of fans a user has. The first term in each of these equations is the rate at which the users see the story. The second terms arise from the rate the story becomes visible in the friends interface of users who are not fans of previous voters but are fans of the most recent voter.

Initially, the story has one vote (from the submitter) and the submitter has S fans, so $v_F(0) = 0$, $v_N(0) = 1$, $F = S$, and $N = U-S-1$ where U is the total number of active users at the time the story is submitted. Over time, a story becomes less visible to users as it moves down the upcoming or (if promoted) front page lists, thereby attracting fewer votes and hence fewer new fans of prior voters.

We use the same visiting rate parameter, ω, for users who are and are not fans of prior voters, since there is only a small correlation between voting activity and the number of fans across all the stories in our dataset, as illustrated in Figure 11. Moreover, many highly active users do not participate in the social network at all (i.e., have neither fans nor friends). Among all users, the correlation between number of votes and number fans is 0.15. More specifically, we assume that with respect to votes on a single story, fans of those voters aren't systematically more likely to visit Digg than other users, such as fans of voters on other stories or users without fans or friends.

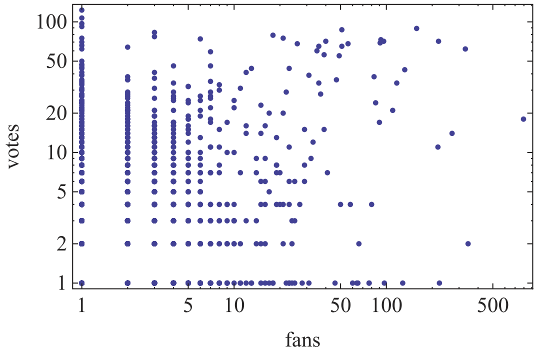


Fig. 11. Comparison of activity (number of votes) and number of fans for each of the 3436 users with at least one vote and one fan.

5.1. Story Visibility

We assume a fan easily sees the story via the friends interface, so $P_F = 1$, as in the previous model [Hogg and Lerman 2009]. Users who are not fans of prior voters must find the story on the front or upcoming pages. Thus P_N depends on how users navigate through these pages and the story's location at the time the user visits Digg. As with the previous model, we use Eq. (3) to describe this behavior.

List position of a story. The page number of a story on the upcoming page q and the front page p at time t is given by Eqs. (5) and (6), with t interpreted as Digg time. The slopes, given in Table III, are the same as with the previous model which averaged over the daily variation in activity. Since each page holds 15 stories, these rates are 1/15th the story submission and promotion rates, respectively.

Since upcoming stories are less popular than the front page, our model has a fraction $c < 1$ of Digg visitors viewing the upcoming stories pages. Combining these effects, we take the visibility of a story at position p in the front page list to be $P_N = f_{\text{page}}(p)$, whereas a story at position q in the upcoming page list is $cf_{\text{page}}(q)$ [Hogg and Lerman 2009].

Promotion to the front page. Promotion to the front page appears to depend mainly on the number of votes the story receives. We model this process by the probability $P(v)$ at which an upcoming story is promoted after its vth vote. We take $P(1) = 0$, that is, a story is not promoted based just on the submitter's vote. The probability a story is not promoted by the time it receives v votes is $\prod_{i=1}^{v}(1 - P(i))$. Stories not promoted are eventually removed, typically 24 hours after submission.

Based on our data, Figure 12 shows the probability $P(v)$ at which an upcoming story is promoted after v votes conditioned on it not having been promoted earlier. We find a significant spread in the number of votes a story has when it is promoted. For predicting whether and when a story will be promoted in our model, we use a logistic regression fit to these values, as shown in the figure. This contrasts with the step function for promotion at 40 votes used in the previous model [Hogg and Lerman 2009].

Friends interface. The fans of the story's submitter can find the story via the friends interface. As additional people vote on the story, their fans can also see the story. We model this with $F(t)$, the number of fans of voters on the story by time t who have not yet seen the story. Although the number of fans is highly variable, we use the average number of additional fans from an extra vote, ρN, in Eq. (12).

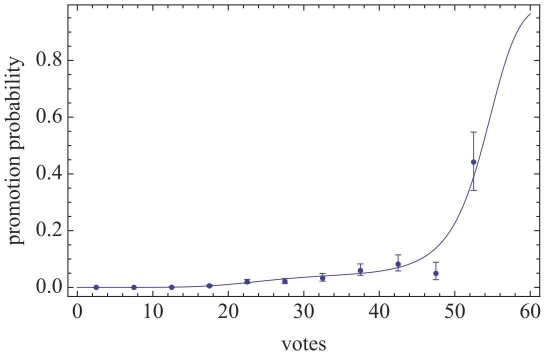


Fig. 12. Probability for promotion before the next vote for an upcoming story as a function of the number of votes. The error bars indicate the 95% confidence intervals for the estimates. The curve is a logistic fit.

5.2. Parameter Estimation

Since we observe votes, not visits to Digg, there is some ambiguity in the rate ω and the interestingness values r_F, r_N. For example, a given value of ωr_F could arise from users often visiting Digg but rarely voting on stories, or less frequent visits with a high chance of voting during each visit. This arbitrary scaling does not affect our focus on the relative behavior of fans and nonfans. For definiteness, we pick a specific value for ω and give interestingness values relative to that choice.

We used the May data to estimate the story location parameters k_{u} and k_{f}. Their values correspond to 54 and 2.7 stories per hour submitted and promoted, respectively.

5.2.1. Estimating Parameters from Observed Votes. In our model, story location affects visibility only for nonfan voters, since fans of prior voters see the story via the friends interface. Thus we just use the nonfan votes to estimate visibility parameters, via maximum likelihood. Specifically, we use the nonfan votes for 16 stories in the June dataset to estimate c and the "law of surfing" parameters μ and λ. We then use fan votes for these stories to evaluate the probability a user is a fan of a new voter, ρ. Separating votes by the different interfaces by which users find stories provides more precise estimation than the prior model [Hogg and Lerman 2009].

This estimation involves comparing the observed votes to the voting rate from the model. As described above, the model uses rate equations to determine the average behavior of the number of votes. A simple approach to relate this average to the observed number of votes is to assume the votes from nonfan users form a Poisson process whose expected value is $dv_N(t)/dt$, given by Eq. (11). This rate changes with time and depends on the model parameters.

For a Poisson process with a constant rate v, the probability to observe n events in time T is the Poisson distribution $e^{-vT}(vT)^n/n!$. This probability depends only on the *number* of events, not the specific times at which they occur. Thus estimating a constant rate involves maximizing this expression, which gives $v = n/T$, that is, the maximum-likelihood estimate of the rate for a constant Poisson process is equal to the average rate of the observed events.

In our case, the voting rate changes with time, requiring a generalization of this estimation. Specifically, consider a Poisson process with nonnegative rate $v(t)$ which depends on one or more parameters to be estimated. Thus, in a small time interval $(t, t + \Delta T)$, the probability for a vote is $v(t)\Delta t$; and this is independent of votes in other time intervals, by the definition of a Poisson process. Suppose we observe n votes at times $0 < t_1 < t_2 \ldots < t_n < T$ during an observation time interval $(0, T)$. Considering small time intervals Δt around each observation, the probability of this observation is

$$\begin{aligned} P(\text{no vote in } (0, t_1))v(t_1)\Delta t &\times \\ P(\text{no vote in } (t_1, t_2))v(t_2)\Delta t &\times \\ &\ldots \\ P(\text{no vote in } (t_{n-1}, t_n))v(t_n)\Delta t &\times \\ P(\text{no vote in } (t_n, T)). \end{aligned}$$

The probability for no vote in the interval (a, b) is

$$\exp\left(-\int_a^b v(t)dt\right).$$

Thus the log-likelihood for the observed sequence of votes is

$$-\int_0^T v(t)dt + \sum_i \log v(t_i).$$

The maximum-likelihood estimation for parameters determining the rate $v(t)$ is a trade-off between these two terms: attempting to minimize $v(t)$ over the range $(0, T)$ to increase the first term while maximizing the values $v(t_i)$ at the specific times of the observed votes. If $v(t)$ is constant, this likelihood expression simplifies to $-vT + n \log v$ with maximum at $v = n/T$, as discussed above for the constant Poisson process. When $v(t)$ varies with time, the maximization selects parameters giving relatively larger $v(t)$ values where the observed votes are clustered in time.

We combine this log-likelihood expression from the votes on several stories, and maximize the combined expression with respect to the story-independent parameters of the model, with the interestingness parameters determined separately for each story.

5.2.2. Estimating Number of Active Users. Our model involves a population of "active users" who visit Digg during our sample period. Specifically, the model uses the rate users visit Digg, ωU. We do not observe visits in our data, but can infer the relevant number of active users, U, from the heterogeneity in the number of votes by users. The June dataset consists of 16283 users who voted at least once during the sample period. Figure 13 shows the distribution of this activity on front page stories. Most users have little activity during the sample period, suggesting that a large fraction of users vote infrequently enough to never have voted during the period data was collected. This behavior can be characterized by an activity rate for each user. A user with activity rate ν will, on average, vote on νT stories during a sample time T. We model the observed votes as arising from a Poisson process whose expected value is νT and the heterogeneity arising from a lognormal distribution of user activity rates [Hogg and Szabo 2009]. This model gives rise to the extended activity distribution while accounting for the discrete nature of the observations. The latter is important for the majority of users who have low activity rates, and so will vote only a few times, or not at all, during our sample period.

Specifically, for n_k users with k votes during the sample period, this mixture of lognormal and Poisson distributions [Bulmer 1974; Miller 2007] gives the log-likelihood of the observations as

$$\sum_k n_k \log P(\mu, \sigma; k),$$

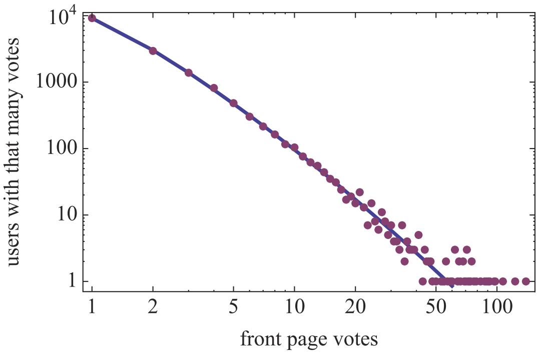


Fig. 13. User activity distribution on logarithmic scales. The curve shows the fit to the model described in the text.

where $P(\mu, \sigma; k)$ is the probability of a Poisson distribution to give k votes when its mean is chosen from a lognormal distribution $P_{\text{lognormal}}$ with parameters μ and σ. From Eq. (9),

$$P(\mu, \sigma; k) = \frac{1}{\sqrt{2\pi}\sigma k!} \int_0^\infty \rho^{k-1} e^{-\frac{(\log(\rho)-\mu)^2}{2\sigma^2} - \rho} d\rho$$

for integer $k \geq 0$. We evaluate this integral numerically. In terms of our model parameters, the value of μ in this distribution equals νT.

Since we don't observe the number of users who did not vote during our sample period, that is, the value of n_0, we cannot maximize this log-likelihood expression directly. Instead, we use a zero-truncated maximum likelihood estimate [Hilbe 2008] to determine the parameters μ and σ for the vote distribution of Figure 13. Specifically, the fit is to the probability of observing k votes conditioned on observing at least one vote. This conditional distribution is $P(\mu, \sigma; k)/(1 - P(\mu, \sigma; 0))$ for $k > 0$, and the corresponding log-likelihood is

$$\sum_{k>0} n_k \log P(\mu, \sigma; k) - U_+ \log(1 - P(\mu, \sigma; 0)),$$

where U_+ is the number of users with at least one vote in our sample, that is, 16283. Maximizing this expression with respect to the distribution's parameters μ and σ gives νT lognormally distributed with the mean and standard deviation of $\log(\nu T)$ equal to -2.06 ± 0.03 and 1.82 ± 0.03, respectively. With these parameters, $P(\mu, \sigma; 0) = 0.757$, indicating about 3/4 of the users had sufficiently low, but nonzero, activity rate that they did not vote during the sample period. We use this value to estimate U, the number of active users during our sample period: $U = U_+/(1 - P(\mu, \sigma; 0))$.

Based on this fit, the curve in Figure 13 shows the expected number of users with each number of votes, that is, the value of $UP(\mu, \sigma; k)$ for $k > 0$. This is a discrete distribution: the lines between the expected values serve only to distinguish the model fit from the points showing the observed values. A bootstrap test [Efron 1979] based on the Kolmogorov–Smirnov (KS) statistic shows the vote counts are consistent with this distribution (p-value 0.48). This test and the others reported in this article account for the fact that we fit the distribution parameters to the data [Clauset et al. 2009].

5.2.3. Estimated Parameters. Table III lists the estimated parameters. We estimate r_F and r_N for each story from its fan and nonfan votes.

The page view distribution seen in this dataset indicates users who choose to visit the upcoming pages tend to explore those pages fairly deeply. This contrasts with the

Table III. Model Parameters, with Times in “Digg Hours”

parameter	value
average rate each user visits Digg	$\omega = 0.2/hr$
number of active users	$U = 70,000$
fraction viewing upcoming pages	$c = 0.065$
page view distribution	$\mu = 6.3$
	$\lambda = 0.14$
probability a user is a voter’s fan	$\rho = 9.48 \times 10^{-6}$
upcoming stories location	$k_{\mathrm{u}} = 3.60\,\mathrm{pages}/hr$
front page location	$k_{\mathrm{f}} = 0.18\,\mathrm{pages}/hr$
story specific parameters	
interestingness to fans	r_F
interestingness to non-fans	r_N
number of submitter’s fans	S

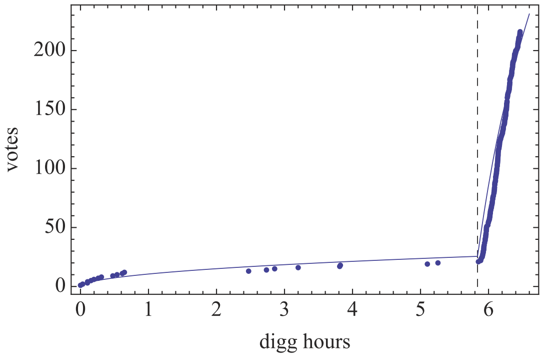


Fig. 14. Voting behavior: the number of votes vs. time, measured in Digg hours, for a promoted story in June 2006. The curve shows the corresponding solution from our model and the dashed vertical line indicates when the story was promoted to the front page. This story eventually received 2566 votes.

more limited exploration, that is, smaller value of μ, seen in the May dataset which included votes well after promotion [Hogg and Lerman 2009]. This suggests differing levels of perseverance of users who visit the upcoming stories compared to the majority of users who focus on front page stories. Alternatively, there could be other ways nonfan users find content that has already moved far down the list of stories.

5.3. Results

Figure 14 compares the solution of the rate equations with the actual votes for one story. The model correctly reproduces the dynamics of voting while the story is on the upcoming stories list and immediately after promotion.

We use the model to evaluate systematic differences in story interestingness between fans and nonfans, with the resulting distribution of values shown in Figure 15. The interestingness values for fans and nonfans of prior voters each have a wide range of values, but the interestingness to fans is generally much higher than to nonfans. Both sets of values fit well to lognormal distributions, as indicated in Figure 16. Specifically, the r_N values fit well to a lognormal distribution with maximum likelihood estimates of the mean and standard deviation of $\log(r_N)$ equal to -4.0 ± 0.1 and 0.63 ± 0.07, respectively, with the ranges giving the 95% confidence intervals. A bootstrap test based on the KS statistic shows the r values are consistent with this distribution (p-value 0.1).

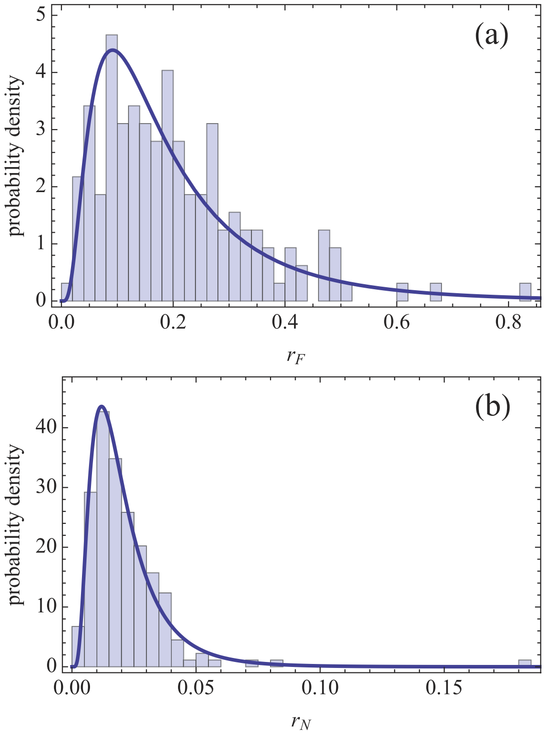


Fig. 15. Distribution of story interestingness for (a) fans and (b) nonfans. The curves are lognormal fits to the values. Note the different ranges for the horizontal scales in the two plots: r_F values tend to be significantly larger than r_N values.

Because there are relatively few votes by fans, we have a larger variance in estimates of r_F than for r_N. In particular, 17 stories have no votes by fans leading to a maximum likelihood estimate $r_F = 0$, though with a large confidence interval. The remaining values are approximately lognormally distributed with maximum likelihood estimates of the mean and standard deviation of $\log(r_F)$ equal to -1.8 ± 0.1 and 0.75 ± 0.08, respectively. The KS statistic indicates the weaker fit, with a p-value of 0.04. Due to the relatively few votes, the discrete nature of the observations likely significantly affects the estimates. For example, a story with no fans among the early votes may reflect a submitter with no fans and a low, but nonzero, interestingness for fans. A subsequent vote by a highly connected user would expose the story to many fans, possibly leading to many votes that the model would miss by assuming $r_F = 0$. One approach to this difficulty is using the lognormal distributions of r values as priors in the estimation. This procedure improves performance somewhat, as discussed below.

Overall, Figure 17 shows there is little relation between how interesting a story is to fans and other users: the correlation between r_F and r_N is -0.11. A randomization test indicates this small correlation is only marginally significant, with p-value 0.05 of arising from uncorrelated values. The relationship between interestingness for fans and other users indicates a considerable variation in how widely stories appeal to the general user community. Specifically, the ratio r_F/r_N ranges from 0 to 87, with median 9.3. The high values correspond to stories that do not get a large number of votes, indicating they are of significantly more interest to the fans of voters than to the general

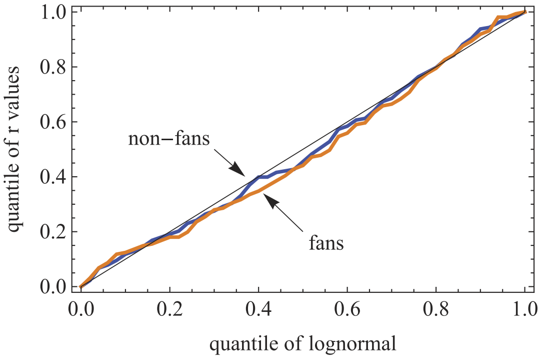


Fig. 16. Quantile-quantile plot comparing the observed distribution for r_F (fans) and r_N (nonfans) with the corresponding lognormal distribution fits (thick curves). For comparison, the thin straight line from 0 to 1 corresponds to a perfect match between the data and the distribution.

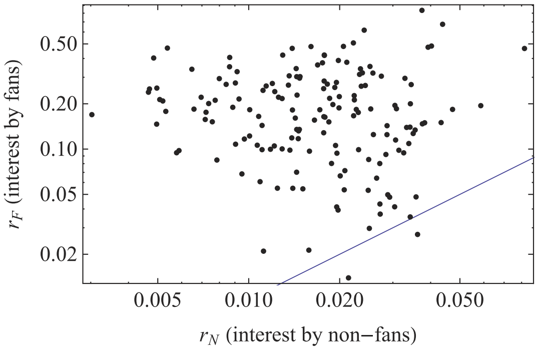


Fig. 17. Log-log plot comparing estimated interestingness to fans (r_F) and nonfans (r_N) for 161 promoted stories with votes from fans (so the estimate of r_F is positive). All the stories in our dataset had nonfan votes, giving all the estimates for r_N as positive numbers. The line indicates where $r_F = r_N$.

user population, that is, “niche interest” stories (corresponding to the upper-left points in Figure 17). As described below, this observation is useful to improve prediction of how popular a story will become based on the reaction of early voters. Identifying niche interest stories could also aid user interface design by selectively highlighting stories on the friends interface that have particularly large estimated values of r_F. Stories with high ratios of r_F/r_N tend to be promoted after fewer votes than those stories with low ratios.

An earlier study [Lerman and Galstyan 2008] noted a curious phenomenon: namely, stories that initially spread quickly through the network, that is, received a large proportion of early votes from fans, ended up not becoming very popular; vice versa, stories that initially spread slowly through the fan network end up becoming popular. This phenomenon appears to be a generic feature of information diffusion on social networks, and has also been observed on blog networks [Colbaugh and Glass 2010] and in Second Life [Bakshy et al. 2009].

Figure 18 shows that our model explains this relationship, which arises from the difference in interestingness for fans and nonfans. Specifically, a low fraction of early votes by fans indicates r_N is relatively large to produce the early nonfan votes in spite of the lower visibility of upcoming stories to nonfan users. Once the story is promoted, it then receives relatively more votes from the general user community (most of whom

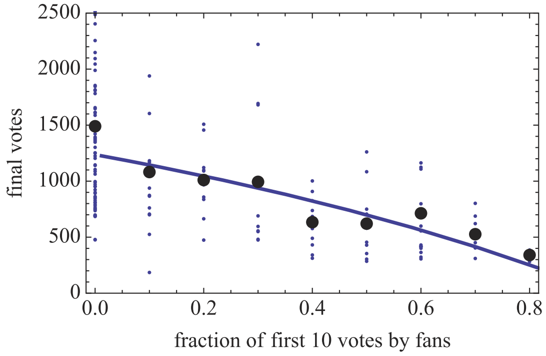


Fig. 18. Relation between final number of votes and the fraction of votes by fans among a story's first 10 votes. Small points are individual stories and the large points are the mean values for each number of votes by fans. The curve shows the model prediction.

are not fans of prior voters). The separation of effects of visibility and interestingness with our model improves this discrimination compared to just using the raw number of votes by fans and nonfans without regard for the story visibility at the time of the votes. For example, the correlation between the final number of votes and r_N/r_F is 0.72 compared to 0.64 for the correlation between the final number of votes and v_N/v_F.

5.4. Discussion

This model with niche interests captures the consequences of link choices: people tend to become fans of users who submit or vote on stories of interest to themselves. The ease of incorporating such additional detail is a useful feature of stochastic models.

Comparing the two models illustrates the practical challenges of incomplete or limited data. For example, data scraped from a Web site can have errors due to unusual user names or unanticipated characters in story titles. Even when Web sites provide an interface to collect data (as Digg provided after the data used in this article was extracted), subtle differences in interpretation of the data fields can still arise, as when users who no longer have a Digg account are all given the same name "inactive," and hence appear to be the same user if not specifically checked for in the script collecting the data.

In particular, the "law of surfing" parameter estimates for the two models are significantly different, a consequence of the log-likelihood being a fairly flat function of these parameters. This arises due to the relatively weak constraints that vote history provides on *views*, that is, how many pages of upcoming or front page stories users choose to view during a visit to Digg. Data including the pages users actually view would provide stronger constraints on this behavior. Such data is not publicly available for Digg. Similarly, the promotion algorithm used by Digg is deliberately not made public to reduce the potential for story submitters to game the system. The stochastic approach identifies potential advantages to Web site providers who have access to more precise data about user behavior, and can therefore make better use of the models, for example, to accurately predict popularity of newly submitted content. More generally, the sensitivity of parameters to the available measured data can suggest additional aspects of user behavior that would be most useful to determine, leading to more focus in future data collection and instrumentation of Web communities. Alternatively, when the models indicate several different types of data could provide the required information, selecting the types most acceptable to the user community (e.g., privacy preserving) can facilitate the data collection while providing opportunities for more accurate models to guide the development of the Web site and its usefulness to its community.

A related data quality issue is the length of time over which data is collected. On the one hand, collecting data for long periods can improve model parameter estimation by providing many more samples. On the other hand, Web sites often rearrange or add features to their user interfaces, which change how users find content. Digg also occasionally changes the promotion algorithm. That is, the stochastic behavior associated with the site is nonstationary rather than arising from a fixed distribution. Moreover, over longer periods of time, new users join the site and some users become inactive. Thus we can't simply improve the model parameter estimation by collecting data over longer periods of time [Hogg and Szabo 2009; Wilkinson 2008]. Instead, the models must be extended to include these additional time-dependent behaviors.

In addition to improving quantitative estimation, similar qualitative behaviors seen with different models identify areas for further investigation. For example, in the two models presented here, the distribution of interestingness over the stories shows a lognormal distribution. This suggests there is an underlying multiplicative process giving rise to the observed values [Mitzenmacher 2004; Redner 1990]. Specifically, the lognormal distribution arises from the multiplication of random variables in the same way that the central limit theorem leads to the normal distribution from the addition of random variables under weak restrictions on their variance and correlations. Thus an important question raised by these models is identifying the story characteristics and user behaviors that combine multiplicatively to lead to the observed lognormal distributions. Identifying such properties would give a more detailed understanding of what leads to interesting content, independent of the effects of visibility provided by the Web site.

6. MODEL-BASED PREDICTION

As discussed in the Introduction, predicting popularity in social media from intrinsic properties of newly submitted content is difficult [Salganik et al. 2006]. However, users' early reactions provide some measure of predictability [Gruhl et al. 2005; Hogg and Szabo 2009; Kaltenbrunner et al. 2007; Lerman and Galstyan 2008; Lerman and Hogg 2010; Szabo and Huberman 2010]. The early votes on a story allow us to estimate its interestingness to fans and other users. We can then use the model to predict how the story will accumulate additional votes. These predictions are for expected values and cannot account for the large variation due, for example, to a subsequent vote by a highly connected user, which leads to a much larger number of votes. However, even with this caveat, we show empirically that we can predict popularity of many stories in our sample.

As one prediction example, we evaluate whether a story will receive at least 500 votes. Predicting whether a story will attract a large number of votes, rather than the precise number of votes, is a useful criterion for predicting whether the story will "go viral" and become very popular. This is exactly Digg's intention behind using crowd sourcing to select a subset of submitted content to feature on the front page [Lerman and Galstyan 2008]. The 500 vote threshold is a useful rule of thumb, as that is close to the median popularity value in a large sample of Digg stories [Lerman and Ghosh 2010; Wu and Huberman 2007].

Table IV compares the predictions with different methods, including a constrained version of our model with $r_F = r_N$, which assumes no systematic difference in interest between fans and other users (similar to the model presented in Section 3). The table gives the prediction error, that is, the fraction of stories misclassified by the prediction algorithm. Misclassified stories include both stories with more than 500 votes

Table IV. Prediction Errors on Whether a Story Receives at least 500 Votes. (The table compares three methods: (1) the full model which allows distinct values for r_F and r_N; (2) the model constrained to have $r_F = r_N$; and (3) direct extrapolation from the rate the story accumulates votes. This comparison involves 178 promoted stories, of which 137 receive at least 500 votes)

	model		direct
	distinct r	same r	extrapolation
first 216 votes	10%	12%	21%
first 10 votes	18%	23%	29%

predicted to be unpopular (false negatives) and unpopular stories predicted to receive more than 500 votes (false positives).

We also compare with direct extrapolation from the early votes. In this procedure, with v votes observed at time t, we extrapolate to vt_{final}/t, where we take t_{final} to be 72 hours since submission, a time by which stories have accumulated all, or nearly all, the votes they will ever get. We use a least-squares linear fit between these observed and extrapolated values. A pairwise bootstrap test indicates the model has a lower prediction error than this extrapolation with p-value of 10^{-2}.

This extrapolation method is similar to that used to predict final votes from the early votes [Szabo and Huberman 2010], but with two differences: (1) we consider a fixed number of votes v (e.g., 10) and, for each story, extrapolate from the time t required for that story to acquire v votes (instead of the opposite choice of considering a fixed prediction time t and, for each story, extrapolating from the number of votes that story has at that time); and (2) we use early votes after submission (i.e., including when the story is upcoming, where the social network has a large effect) instead of early votes after promotion.

In the case of prediction based on the first 10 votes, which is before the stories are promoted, an additional question is how well the model predicts whether the story will eventually be promoted. We find a 25% error rate in predicting promotion based on the first 10 votes.

We can improve predictions from early votes by using the lognormal distributions of r_F and r_N, shown in Figure 15, as the prior probability to combine with the likelihood from the observations according to Bayes theorem. Specifically, instead of maximizing the likelihood of the observed votes, $P(r|\text{votes})$, as discussed above, this approaches maximizes the posterior probability, which is proportional to $P(r|\text{votes})P_{\mathrm{prior}}(r)$ where P_{prior} is taken to be the lognormal distribution $P_{\mathrm{lognormal}}$ in Eq. (9) with parameters from the fits shown in Figure 15.

This method gives little change in estimates of r_N, due to the relatively large number of nonfan votes on each story. However, using the prior makes large changes in some of the r_F estimates, thereby avoiding the small number of extreme predictions made by poor estimates. Using this prior to aid estimation is particularly significant when there are no votes by fans among the early votes, leading to an estimate of $r_F = 0$, but later a user with many fans votes on the story. In this case, as illustrated in Figure 19, using the lognormal as a prior gives a positive estimate for r_F, thereby predicting some votes by any subsequent users who are fans of earlier voters.

By avoiding these extreme cases, this procedure improves the correlation between predicted and actual final votes, as well as the predicted rank ordering of the stories (i.e., relative popularity of stories) as seen with a larger value of the Spearman rank correlation when using the prior distribution. For example, when predicting based on the first 10 votes, using this prior increases the Spearman rank correlation between predicted and actual number of votes from 0.46 to 0.53. For comparison,

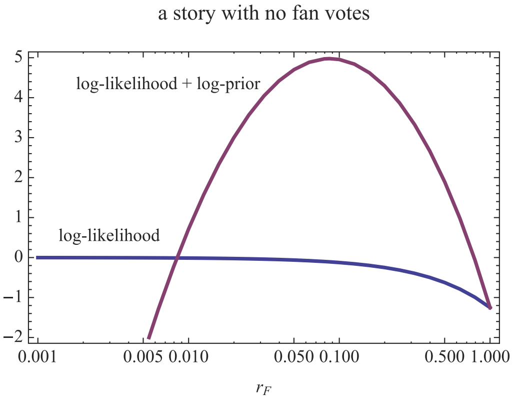


Fig. 19. Comparison of log-likelihood (i.e., log $P(r|\text{votes})$) and log-likelihood plus $\log(P_{\text{prior}}(r))$ for estimating r_F for a story with no fan votes. The maximum of the log-likelihood is at $r_F = 0$ while the maximum with the prior is $r_F = 0.086$.

this correlation for direct extrapolation from the first 10 votes is 0.32, and it is 0.34 for the model constrained to have $r_F = r_N$. Pairwise bootstrap tests indicate the differences between these values are significant with p-values less than 10^{-3}, except the difference between the last two cases has a p-value of 10^{-2}.

Our earlier graph-based method [Lerman and Galstyan 2008] used the fraction of early fan votes to predict whether a newly submitted Digg story will go on to receive more than 505 votes. That method looked at the number of fan votes among the first 10 votes on 39 stories submitted by users who were among the 100 top-ranked Digg users in June 2006. Of these stories, 13 were eventually promoted, but only four went on to receive more than 505 votes. The prediction error of this method is 31%. Looking at the prediction error of the 13 promoted stories is also 31%, slightly worse than direct extrapolation in Table IV. One limitation of this method is that it can only predict popularity of stories submitted by well-connected users: for other stories, users who are not fans of the submitter dominate the votes. Stochastic models have no such limitation and lead to more accurate predictions than those made by the graph-based model.

7. RELATED WORK

The social Web provides massive quantities of data about the behavior of large groups of people online. Researchers are using this data to study a variety of topics, including detecting [Adar et al. 2004; Leskovec et al. 2007] and influencing [Domingos and Richardson 2001; Kempe et al. 2003] trends in public opinion, and dynamics of information flow in groups [Lerman and Ghosh 2010; Leskovec et al. 2007; Wu et al. 2004].

Stochastic modeling offers a flexible mathematical framework for relating details of individual behavior to the collective dynamics of an ensemble of similar individuals. This framework has been used in a variety of disciplines: for example, to understand population dynamics [Haberman 1987], the spread of disease [Bailey 1975; Hethcote 2000] and computer viruses [Kephart and White 1991], to model the flow of traffic [Helbing 2001], crowds of pedestrians [Moussaïd et al. 2011], as well as the behavior of social insects [Franks et al. 2002] and multirobot teams [Agassounon et al. 2004; Lerman and Galstyan 2002; Lerman et al. 2001, 2005]. The stochastic modeling approach described here can apply to any social media site by matching the state diagram (e.g., Figures 9 and 10) to the information on users and content displayed by the site. For example, this approach has been used to model user behavior on a political discussion Web site (Essembly.com) where users propose and discuss topics of current political interest [Brzozowski et al. 2008]. This site differs from Digg in not having an equivalent of Digg's front page, so topics change their visibility more gradually than on Digg. Moreover, the site provides more variety in the types of links users can form, which allows users to separate social contacts from those they do not know personally but whose political views they find especially significant. Thus the details of topic visibility differ from those of Digg. Nevertheless, the stochastic modeling approach applies to this site, and shows similar behaviors among users and the distribution of interestingness among topics [Hogg and Szabo 2009]. Stochastic models have also been applied to the behavior of posts and comments on blogs [Gotz et al. 2009].

Several researchers examined the role of social dynamics in explaining and predicting distribution of popularity of online content. Wilkinson [2008] found broad distributions of popularity and user activity on many social media sites, and showed that these distributions can arise from simple macroscopic dynamical rules. Wu and Huberman [2007] constructed a phenomenological model of the dynamics of collective attention on Digg. Their model is parameterized by a single variable that characterizes the rate of the decay of interest in a news article. Rather than characterize evolution of votes received by a single story, they show the model describes the distribution of final votes received by promoted stories. Our model offers an alternative explanation for the distribution of votes. Rather than novelty decay, we argue that the distribution can also be explained by the combination of a nonuniform variations in the stories' inherent interest to users and the effects of the user interface, specifically decay in visibility as the story moves to subsequent front pages. Such a mechanism can also explain the distribution of popularity of photos on Flickr, which would be difficult to characterize by novelty decay. Crane and Sornette [2008] analyzed a large number of videos posted on YouTube and found that collective dynamics was linked to the inherent quality of videos. By looking at how the observed number of votes received by videos changed in time, they could separate high-quality videos, whether they were selected by YouTube editors or spontaneously became popular, from junk videos. This study is similar in spirit to our own in exploiting the link between observed popularity and content quality. However, while this, and the Wu and Huberman study, aggregated data from tens of thousands of individuals, our method focuses instead on the *microscopic* dynamics, modeling how individual behavior contributes to the observed popularity of content. In Lerman and Hogg [2010] we used the simple model of social dynamics, reviewed in this article, to predict whether Digg stories will become popular. The current article improves on that work.

Researchers found statistically significant correlation between early and late popularity of content on Slashdot [Kaltenbrunner et al. 2007], Digg and YouTube [Szabo and Huberman 2010]. Specifically, similar to our study, Szabo and Huberman [2010] predicted long-term popularity of stories on Digg. Through large-scale statistical study of stories promoted to the front page, they were able to predict the stories' popularity after 30 days based on their popularity one hour after promotion. Unlike our work, their study did not specify a mechanism for evolution of popularity, and simply exploited the correlation between early and late story popularity to make the prediction. Our work also differs in that we predict popularity of stories shortly after submission, long before they are promoted.

Several researchers [Bakshy et al. 2009; Colbaugh and Glass 2010; Lerman and Galstyan 2008] found that early diffusion of information across an interlinked community is a useful predictor of how far it will spread across the network in general. Both Lerman and Galstyan [2008] and Colbaugh and Glass [2010] exploited the anti-correlation between these phenomena to predict final popularity. Specifically, the former work used anti-correlation between the number of early fan votes and the stories' eventual popularity on Digg. Specifically, they found that stories that initially received few votes from the fans of submitters and previous voters went on to become much more popular than stories which had many initial votes from fans. Using this correlation, they were able to predict whether stories submitted by well-connected users will become popular. That work exploited social influence only to make the prediction, and the results were not applicable to stories submitted by poorly connected users which were not quickly discovered by highly connected users. In contrast, the approach described in this article considers effects of social influence regardless of the connectedness of the submitter, and also accounts for story quality in making a prediction about story popularity. More generally, in Digg the visibility of stories to nonfans increases significantly and abruptly upon promotion to the front page, leading to most votes coming from users who are not part of the submitter's social network. Thus Digg provides limited scope for evaluating graph-based methods. Other studies of graph-based methods for prediction [Bakshy et al. 2009; Colbaugh and Glass 2010] thus focus on other settings where social connections have a larger role in the aggregate behavior.

8. CONCLUSION

In the vast stream of new user-generated content, only a few items will prove to be popular, attracting a lion's share of attention, while the rest languish in obscurity. Predicting which items will become popular is exceedingly difficult, even for people with significant expertise. This prediction difficulty arises because popularity is weakly related to inherent content quality, and social influence leads to an uneven distribution of popularity that is sensitive to the early choices of users in the social network. We described how stochastic models of user behavior on a social media Web site can partially address this prediction challenge by quantitatively characterizing evolution of popularity. The model shows how popularity is affected by item quality and social influence. We evaluated the usefulness of this approach for the social news aggregator Digg, which allows users to submit and vote on news stories. The number of votes a story accumulates on Digg shows its popularity. In earlier work we developed a model of social voting on Digg, which describes how the number of votes received by a story changes in time. In that model, knowing how interesting a story is to the user community, on average, and how connected the submitter is fully determines the evolution of the story's votes. This leads to an insight that a model can be used to predict a story's popularity from the initial reaction of users to it. Specifically, we use observations of evolution of the number of votes received by a story shortly after submission to estimate how interesting it is, and then use the model to predict how many votes the story will get after a period of a few days. Model-based prediction outperforms other methods that exploit social influence only, and also correlation between early and late votes received by stories. We improved prediction by developing a more fine-grained model that differentiates between how interesting a story is to fans and to the general population.

These results demonstrate the applicability of the stochastic approach to social media, in spite of the large variations in user participation and interestingness of the content. One interesting open question is the nature of the social influence on user behavior. In our model, the influence has two components: increased visibility of a story to fans due to the friends interface and the higher interestingness of the story to fans. This higher interestingness could be due to self-selection, whereby users become fans of people whose submissions or votes are of particular interest. Alternatively, users could be directly influenced by the activities of others [Salganik et al. 2006], with the possibility that this influence depends not just on whether friends vote on a story but also how many friends do so [Centola 2010]. Other challenging open questions include identifying common mechanisms underlying the observed regularities, accounting for time-dependent changes in the Web site and the user community, and extending the approach to a wider variety of Web sites.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the Fetch Corporation for providing data extraction tools.

REFERENCES

ADAR, E., ZHANG, L., ADAMIC, L. A., AND LUKOSE, R. M. 2004. Implicit structure and the dynamics of blogspace. In *Proceedings of the Workshop on the Weblogging Ecosystem, 13th International World Wide Web Conference*.

AGARWAL, N., LIU, H., TANG, L., AND YU, P. S. 2008. Identifying the influential bloggers in a community. In *Proceedings of the International Conference on Web Search and Web Data Mining (WSDM'08)*. ACM, New York, 207–218.

AGASSOUNON, W., MARTINOLI, A., AND EASTON, K. 2004. Macroscopic modeling of aggregation experiments using embodied agents in teams of constant and time-varying sizes. *Auton. Robots 17*, 2–3, 163–191.

ANDERSON, C. 2006. *The Long Tail: Why the Future of Business is Selling Less of More*. Hyperion.

BAILEY, N. 1975. *The Mathematical Theory of Infectious Diseases and its Applications*. Griffin, London.

BAKSHY, E., KARRER, B., AND ADAMIC, L. A. 2009. Social influence and the diffusion of user-created content. In *Proceedings of the 10th ACM Conference on Electronic Commerce (EC'09)*. 325–334.

BARABÁSI, A.-L. 2005. The origin of bursts and heavy tails in human dynamics. *Nature 435*, 207–211.

BRZOZOWSKI, M. J., HOGG, T., AND SZABO, G. 2008. Friends and foes: Ideological social networking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing (CHI'08)*. ACM, New York, 817–820.

BULMER, M. G. 1974. On fitting the Poisson lognormal distribution to species-abundance data. *Biometrics 30*, 101–110.

CENTOLA, D. 2010. The spread of behavior in an online social network experiment. *Science 329*, 1194–1197.

CLAUSET, A., SHALIZI, C. R., AND NEWMAN, M. E. J. 2009. Power-law distributions in empirical data. *SIAM Rev. 51*, 661–703.

COLBAUGH, R. AND GLASS, K. 2010. Early warning analysis for social diffusion events. In *Proceedings of the IEEE International Conferences on Intelligence and Security Informatics*.

CRANE, R. AND SORNETTE, D. 2008. Viral, quality, and junk videos on YouTube: Separating content from noise in an information-rich environment. In *Proceedings of the AAAI Symposium on Social Information Processing*. AAAI, Menlo Park, CA.

DOMINGOS, P. AND RICHARDSON, M. 2001. Mining the network value of customers. In *Proceedings of the KDD*.

EFRON, B. 1979. Bootstrap methods: Another look at the jackknife. *Annals of Statistics 7*, 1–26.

ELLNER, S. AND GUCKENHEIMER, J. 2006. *Dynamic Models in Biology*. Princeton University Press, Princeton, NJ.

FRANKS, N. R., PRATT, S. C., MALLON, E. B., BRITTON, N. F., AND SUMPTER, D. J. T. 2002. Information flow, opinion polling and collective intelligence in house-hunting social insects. *Philosophical Trans. Royal Soc. London B, Biol. Sci. 357*, 1427, 1567–1583.

GÓMEZ, V., KALTENBRUNNER, A., AND LÓPEZ, V. 2008. Statistical analysis of the social network and discussion threads in Slashdot. In *Proceedings of the 17th International Conference on World Wide Web (WWW'08)*. ACM, New York, 645–654.

GOTZ, M., LESKOVEC, J., MCGLOHON, M., AND FALOUTSOS, C. 2009. Modeling blog dynamics. In *Proceedings of the 3rd International Conference on Weblogs and Social Media (ICWSM'09)*. AAAI, 26–33.

GRUHL, D., GUHA, R., KUMAR, R., NOVAK, J., AND TOMKINS, A. 2005. The predictive power of online chatter. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD'05)*. ACM, New York, 78–87.

HABERMAN, R. 1987. *Mathematical Models: Mechanical Vibrations, Population Dynamics, and Traffic Flow*. SIAM.

HELBING, D. 2001. Traffic and related self-driven many-particle systems. *Rev. Modern Phys. 73*, 4, 1067–1141.

HETHCOTE, H. W. 2000. The mathematics of infectious diseases. *SIAM Rev. 42*, 4, 599–653.

HILBE, J. M. 2008. *Negative Binomial Regression*. Cambridge University Press.

HOGG, T. AND LERMAN, K. 2009. Stochastic models of user-contributory web sites. In *Proceedings of the 3rd International Conference on Weblogs and Social Media (ICWSM'09)*. AAAI, 50–57.

HOGG, T. AND SZABO, G. 2009. Diversity of user activity and content quality in online communities. In *Proceedings of the 3rd International Conference on Weblogs and Social Media (ICWSM'09)*. AAAI, 58–65.

HUBERMAN, B. A., PIROLLI, P. L. T., PITKOW, J. E., AND LUKOSE, R. M. 1998. Strong regularities in World Wide Web surfing. *Science 280*, 95–97.

KALTENBRUNNER, A., GOMEZ, V., AND LOPEZ, V. 2007. Description and prediction of Slashdot activity. In *Proceedings of the 5th Latin American Web Congress (LA-WEB'07)*.

KAMPEN, N. G. V. 1992. *Stochastic Processes in Physics and Chemistry*. Elsevier Science, Amsterdam.

KEMPE, D., KLEINBERG, J., AND ÉVA TARDOS. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'03)*. ACM, New York, 137–146.

KEPHART, J. O. AND WHITE, S. R. 1991. Directed-graph epidemiological models of computer viruses. In *Proceedings of the IEEE Symposium on Security and Privacy*. IEEE, Los Alamitos, CA.

KITTUR, A., CHI, E., PENDLETON, B. A., SUH, B., AND MYTKOWICZ, T. 2006. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. In *Proceedings of the World Wide Web Conference*.

LERMAN, K. 2007a. Social information processing in social news aggregation. *IEEE Internet Comput.* (Special Issue on Social Search) *11*, 6, 16–28.

LERMAN, K. 2007b. Social networks and social information filtering on Digg. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM'07)*.

LERMAN, K. AND GALSTYAN, A. 2002. Mathematical model of foraging in a group of robots: Effect of interference. *Autonom. Robots 13*, 2, 127–141.

LERMAN, K. AND GALSTYAN, A. 2008. Analysis of social voting patterns on digg. In *Proceedings of the 1st ACM SIGCOMM Workshop on Online Social Networks*. ACM, New York.

LERMAN, K. AND GHOSH, R. 2010. Information contagion: An empirical study of spread of news on Digg and Twitter social networks. In *Proceedings of the 4th International Conference on Weblogs and Social Media (ICWSM)*.

LERMAN, K. AND HOGG, T. 2010. Using a model of social dynamics to predict popularity of news. In *Proceedings of the 19th International World Wide Web Conference (WWW)*.

LERMAN, K. AND JONES, L. 2007. Social browsing on Flickr. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM'07)*.

LERMAN, K., GALSTYAN, A., MARTINOLI, A., AND IJSPEERT, A. 2001. A macroscopic analytical model of collaboration in distributed robotic systems. *Artif. Life J. 7*, 4, 375–393.

LERMAN, K., MARTINOLI, A., AND GALSTYAN, A. 2005. A review of probabilistic macroscopic models for swarm robotic systems. In *Swarm Robotics Workshop: State-of-the-Art Survey*, Lecture Notes in Computer Science, vol. 3342, Springer, Berlin, 143–152.

LESKOVEC, J., ADAMIC, L., AND HUBERMAN, B. 2007. The dynamics of viral marketing. *ACM Trans. Web 1*, 1.

LESKOVEC, J., KRAUSE, A., GUESTRIN, C., FALOUTSOS, C., VANBRIESEN, J., AND GLANCE, N. 2007. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'07)*. ACM, New York, 420–429.

LLOYD-SMITH, J., SCHREIBER, P., KOPP, P., AND GETZ, W. 2005. Superspreading and the effect of individual variation on disease emergence. *Nature 438*, 355–359.

MILLER, G. 2007. Statistical modeling of Poisson/log-normal data. *Radiation Protection Dosimetry 124*, 155–163.

MITZENMACHER, M. 2004. A brief history of generative models for power law and lognormal distributions. *Internet Math. 1*, 226–251.

MORENO, Y., PASTOR-SATORRAS, R., AND VESPIGNANI, A. 2002. Epidemic outbreaks in complex heterogeneous networks. *Euro. Phys. J. B Condens. Matter Complex Syst. 26*, 4, 521–529.

MOUSSAÏD, M., HELBING, D., AND THERAULAZ, G. 2011. How simple rules determine pedestrian behavior and crowd disasters. *Proc. Nat. Acad. Sci. 108*, 17, 6884–6888.

OPPER, M. AND SAAD, D. 2001. *Advanced Mean Field Methods: Theory and Practice*. MIT Press, Cambridge, MA.

REDNER, S. 1990. Random multiplicative processes: An elementary tutorial. *Am. J. Phys. 58*, 267–273.

SALGANIK, M., DODDS, P., AND WATTS, D. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science 311*, 854.

SORNETTE, D. 2004. *Critical Phenomena in Natural Sciences: Chaos, Fractals, Self-organization and Disorder: Concepts and Tools* 2nd Ed. Springer, Berlin.

SZABO, G. AND HUBERMAN, B. A. 2010. Predicting the popularity of online content. *Comm. ACM 53*, 8, 80–88.

VÁZQUEZ, A., OLIVEIRA, J. G., DEZSÖ, Z., GOH, K., KONDOR, I., AND BARABÁSI, A. 2006. Modeling bursts and heavy tails in human dynamics. *Phys. Rev. E 73*, 3, 036127+.

WILKINSON, D. M. 2008. Strong regularities in online peer production. In *Proceedings of the 9th ACM Conference on Electronic Commerce (EC'08)*. ACM, New York, 302–309.

WU, F. AND HUBERMAN, B. A. 2007. Novelty and collective attention. *Proc. Nat. Acad. Sci. 104*, 45, 17599–17601.

WU, F., HUBERMAN, B. A., ADAMIC, L. A., AND TYLER, J. R. 2004. Information flow in social groups. *Physica A: Stat. Theor. Phys. 337*, 1–2, 327–335.

Received October 2010; revised July 2011; accepted August 2011